Building Footprint Extraction from VHR Remote Sensing Images Combined with Normalized DSMs using Fused Fully Convolutional Networks

Ksenia Bittner, Fathalrahman Adam, Shiyong Cui, Marco Körner, *Member, IEEE,* Peter Reinartz, *Member, IEEE*

Abstract—Automatic building extraction and delineation from high-resolution satellite imagery is an important but very challenging task, due to the extremely large diversity of building appearances. Nowadays, it is possible to use multiple high-resolution remote sensing data sources which allow the integration of different information in order to improve the extraction accuracy of building outlines. Many algorithms are built on spectral-based or appearance-based criteria, from single or fused data sources, to perform the building footprint extraction. But the features for these algorithms are usually manually extracted, which limits their accuracy. Recently developed *fully convolutional networks (FCNs)*, which are similar to normal *convolutional neural networks (CNNs)*, but the last fully connected layer is replaced by another convolution layer with a large "receptive field", quickly became the state-of-the-art method for image recognition tasks, as they bring the possibility to perform dense pixel-wise classification of input images. Based on these advantages, *i.e.*, the automatic extraction of relevant features, and dense classification of images, we propose an *end-to-end fully convolutional network (FCN)* which effectively combines the spectral and height information from different data sources and automatically generates a full resolution binary building mask. Our architecture (FUSED-FCN4S) consists of three parallel networks merged at a late stage, which helps propagating fine detailed information from earlier layers to higher-levels, in order to produce an output with more accurate building outlines. The inputs to the proposed Fused-FCN4s are *three-band (RGB)*, *panchromatic (PAN)*, and *normalized digital surface model (nDSM)* images. Experimental results demonstrate that the fusion of several networks is able to achieve excellent results on complex data. Moreover, the developed model was successfully applied to different cities to show its generalization capacity.

Index Terms—deep learning, fully convolutional networks, building footprint, binary classification, data fusion, satellite images

1 INTRODUCTION

Since the launch of the first satellite for Earth monitoring, the development of different sensors significantly increased the availability of high-resolution remote sensing imagery, providing a huge potential for meaningful and accurate terrestrial scene interpretation. The analysis of satellite imagery involves the identification of building rooftops as one of the most challenging, but important objects among various terrestrial targets in an image. This information is useful for many remote sensing applications, such as urban planning and reconstruction, disaster monitoring, 3D city modeling, *etc*. A vast amount of manual work is done on interpretation and identification of targets in remote sensing imagery by human interpreters. However, it is very time-consuming and expensive to distinguish buildings from other objects and delineate their contours manually. Therefore, there was a great number of attempts to develop methodologies to extract buildings automatically.

Some algorithms for building detection on the basis of aerial [1] and high-resolution satellite imagery [2, 3] utilize specific criteria of building appearance like the uniform spectral reflectance values [4, 5]. The main problem to be encountered in these approaches is the confusion of the building with other objects with similar spectral reflectance. Many automatic building extraction methods from multi-spectral imagery or *digital surface models (DSMs)*, providing height information for a scene, define the criteria such as the shapes of relatively homogeneous buildings follow a certain pattern [6–8]. However, these methodologies are very limited, because the defined criteria work only for certain types of buildings but fail to generalize to areas with complex and heterogeneous buildings. Different data sources can provide complementary information to each other. As a result, the integration of different data sources creates the opportunity for improving accuracy and robustness of the extraction results. Therefore, recently developed methodologies apply the use of fusing data sources, such as multi-spectral images with either stereo DSM or *light detection and ranging (LIDAR)* DSM rather than the use of only a single data source [9, 10]. Although many approaches have been proposed for building footprint extraction, this topic remains a complex problem for scientists.

With the revolutionary development of deep learning techniques, the definition of task-specific features is not under demand anymore for learning-based image analysis tasks. Instead, the most suitable features can be discovered automatically during the training procedure on a big dataset

- *K. Bittner, F. Adam, S. Cui and P. Reinartz are with the Earth Observation Center of the German Aerospace Center (DLR) Münchner Str. 20, Oberpfaffenhofen, Germany.*
  E-mail: (Ksenia.Bittner,Fathalrahman.Adam, Peter.Reinartz)@dlr.de E-mail: shiyong.cui84@gmail.com
- *M. Körner is with Technical University of Munich, Department of Civil, Geo and Environmental Engineering, Arcisstr. 21, 80333 München, Germany E-mail: marco.koerner@tum.de*

through the organization of multi-layer neural networks. *Convolutional neural networks (CNNs)* [11, 12] are one of the most successful deep learning architectures. They achieved state-of-the-art results and became the dominant approach for image understanding in computer vision. The main objective of this work is to adapt the CNNs for remote sensing imagery understanding with high accuracy. This is a challenging task since the satellite imagery is very different from usual computer vision images in a sense of size, perspective view and semantic meaning of every pixel within the whole scene.

In this paper, we analyze the potential of end-to-end CNN learning and apply it to a dense pixel-wise binary classification problem of building vs. non-building identification. In order to take advantage of multiple remote sensing data, we design a hybrid *fully convolutional network (FCN)* architecture, based on approach [13], to produce dense binary classification maps from raw images. The network performs a late fusion of the pre-trained model derived from ImageNet data for spectral images (RGB and PAN) with DSM features trained from scratch. Besides, the network is augmented with additional connections which provide the top classification layers with the access to high-frequency information and, as a result, makes it possible to predict at a finer spatial resolution. Moreover, we compare the proposed framework with "naïve" fusion of a triple-stream architecture which naïvely averages the predictions from multi-source data and show that the proposed merged neural network improves the prediction accuracy.

Code is available at https://gitlab.com/ksenia_bittner/fused-fcn4s.

The remainder of the paper is arranged as follows. In Section 2, related work for building extraction from earlier approaches to more advanced using CNNs is summarized. The background of CNNs, their transformation to FCNs, and details of our deep network architecture are described in Section 3. In Section 4, we introduce the dataset and present implementation details and training strategies. The experimental results on two different datasets applying the proposed deep network architecture, together with their quantitative evaluation are shown and discussed in Section 5. Section 6 concludes the paper.

2 Related work

A significant amount of work has been done on building extraction from remote sensing imagery. In general, the existing methods can be grouped into two classes according to the information used for building extraction: Aerial or high-resolution satellite imagery and 3D information in the form of DSMs. The earlier studies introduce methodologies based on low-level feature extraction—like edges, line segments and corners—which were grouped together to form building hypotheses [14–17].

It was observed that building rooftops within relatively homogeneous areas have more regular shapes represented by rectangles or combinations of them. As a result, the methodologies employing the shape information were developed. Karantzalos *et al.* [18] integrate multiple shape priors into the segmentation process, for extracting the building footprints from a PAN image. Sirmacek *et al.* [19] extract building boundaries from DSM data based on building skeletons, which are split into various pieces and introduced to a box-fitting algorithm. Then, the active rectangular shape growing is performed, until the difference between the previously extracted building edges and the rectangle is reduced. Guercke *et al.* [16] first detect building edges and separate them from other above-ground information using DSM data and *normalized difference vegetation index (NDVI)*, then iteratively fit a rectangle to the building contour until all building parts become rectangles. Although the algorithms based on geometrical primitives achieve good results, they experience difficulties especially with more complex, non-rectangular building shapes.

Geometrical information, like shape, is a very useful feature for segmentation of remotely sensed images. For example, the shadow information can serve as hints for building location [4, 5] or prediction of its shape and height properties [15]. Moreover, the spectral information also presents another useful data source. The NDVI data extracted from *red* and *near-infrared (NIR)* channels of a multi-spectral image indicate vegetation and, as a result, can help to eliminate trees. The early approaches for image classification typically employ task-specific features like color histograms or local binary patterns and pass them to machine learning algorithms to generate a labeled image [20–22]. Ngo *et al.* [23] decompose an image into small homogeneous regions, which are then grouped into clusters. The assumption that buildings are typically accompanied with shadows is used to merge these building segments with their neighboring regions in the same cluster to produce final building proposals. But the features can be extracted not only from spectral images.

In recent years, data fusion has received significant attention not only in remote sensing but also in many other domains. Its applications include medical and industrial robotics, where pattern-recognition and inference techniques are used to perform tasks ranging from 3D object recognition, to determination of object orientation and localization [24–27], human action recognition [28], surveillance systems designed to detect, track, and identify targets and events [29], autonomous driving [30, 31], *etc*.

In remote sensing, the increasing number of air- and space-borne sensors also led to the emergence of several mixed datasets [32]. The combination of imagery and DSMs is the most prominent application for data fusion, as both modalities have their advantages and limitations. Their integration can help to improve building extraction accuracy, as well as robustness. Sohn *et al.* [9] first identify the isolated building objects by investigating the height property of laser points and NDVI from IKONOS imagery. Then, a full description of building outlines is accomplished by merging convex polygons obtained from the hierarchical division of proposed building region by rectilinear lines using the *binary space partitioning (BSP)* tree. Rottensteiner *et al.* [33] fuse features extracted from the *normalized digital surface model (nDSM)* and RGB images using the Dempster-Shafer methodology [34]. Zabuawala *et al.* [35] extract the initial building footprint, based on an iterative morphological filtering approach. This initial segmentation result is enhanced afterward with color aerial imagery by first generating a combined gradient surface and then applying the watershed

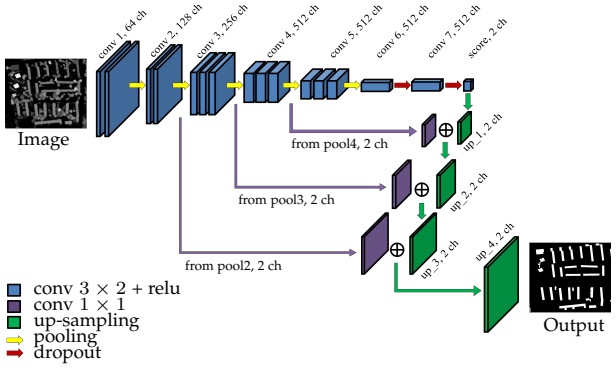


Fig. 1: Schematic representation of our FCN4s architecture.

algorithm to find ridge lines on the surface. Turlapaty *et al.* [36] first obtain an initial test dataset by thresholding those samples from DSM that certainly do not correspond to buildings. Then, the block-based features are extracted from the potential building segments. Finally, these features are used for *support vector machine (SVM)* classification to discriminate buildings from non-building objects in the initial test dataset. Although the methodologies based on hand-crafted features have shown promising results, their main drawback is that they are not robust to the natural large variety of shapes and appearances of buildings within remote sensing images of different scales.

With a tremendous jump in development in the field of artificial neural networks, it became possible to learn image features automatically instead of extracting them by classical methods. A pioneering work in learning large-extent spatial contextual features for labeling an aerial image is Mnih [37]. It utilizes a specific patch-based architecture, where instead of the inference of a single value to classify a whole image, a dense classification patch is retrieved as a final outcome. In order to enhance the performance of the proposed algorithm, the results were processed by *conditional random fields (CRFs)*, as this approach improves the predictions by encouraging smoothness between similar adjacent pixels. However, due to cropping the images to a fixed size, the procedure introduces discontinuities on the border of the classified patches. In our earlier work [38], we present a four-layer *fully connected (FC)* neural network for building footprint extraction from nDSMs. This approach is able to extract the complete building footprints to a high degree of accuracy. But the computation of such network is heavily influenced by the FC layers and the level of details, which directly depends on the patch size.

Since Krizhevsky *et al.* [11] introduced the innovative architecture based on earlier works on deep CNNs [39, 40], they became the state-of-the-art for image recognition tasks. Although CNNs are well established for image classification problems, the methodologies related to segmentation tasks are still under exploration. Socher *et al.* [41] introduce a model based on convolution and pooling layers. In other words, the low-level features learned from CNN layer are given as inputs to multiple *recursive neural networks (RNNs)* in order to build higher-order features. Like in the present paper, this work uses, additionally to RGB, a depth image which is processed in a separate stream. However, in contrast to our work, there is no end-to-end training. Farabet *et al.* [42] assign patch-wise predictions from a CNN with three convolutional layers and a fully connected layer to superpixels which are combined into meaningful regions after applying a CRF. Similar to our work, that approach processes each scale from the generated image pyramids separately with the CNN but the filter weights are shared across scales.

In the field of semantic segmentation, it became more popular to follow the idea of FCNs proposed by Long *et al.* [13]. The FCNs are the type of CNNs which consist of convolutional and pooling layers plus activation functions. Thus, there are no FCs layers in this type of network. As a result, they can compute spatially explicit label maps efficiently and are independent from input size. To deal with the loss of spatial resolution due to the pooling layers or filters applied not on every pixel but skipping some convolutions through, the series of papers propose to up-sample the probability maps back to the resolution of the input image. A similar approach to ours for recovering high-frequency information is presented in the *U-Net* architecture [43]. Each step of the upper part of the network is comprised of 2×2 convolutions ("up-convolution") concatenated with the correspondingly cropped feature maps from the lower part of the network and 3×3 convolutions. The final layer is a 1×1 convolution which brings the number of layers in the last layer to the desired number of classes. In contrast, fully convolutional *DenseNet* [44] approach recovers higher frequencies by using a so-called *Transition Up* block. This block is composed of a transposed convolution to up-sample the incoming feature map, then a skip connection is used to concatenate the input of the Transition Up block with the up-sampled features, producing the final output of the block at the target resolution. In the context of building footprint extraction, Yuan [45] proposes a type of FCN architecture where the outputs of each stage of the network are up-sampled, stacked together, and fed into a convolutional layer with a filter of size $1 \times 1 \times n$ (where n is the number of stacked feature maps). A sort of prediction map is generated, where, in contrast to our

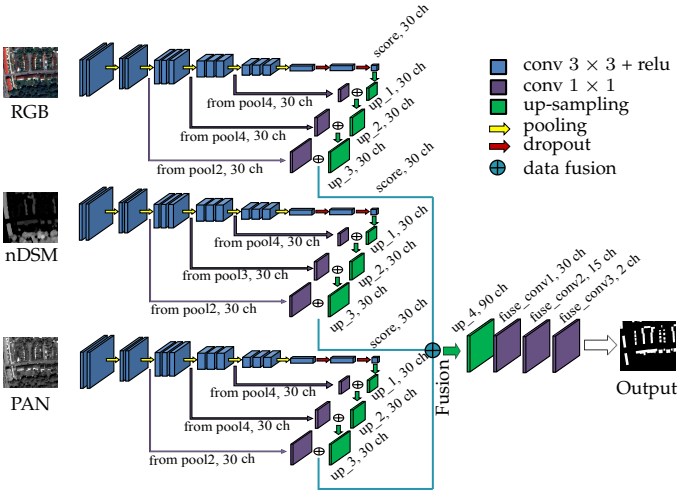


Fig. 2: Schematic representation of the proposed Fused-FCN4s architecture.

work, the values of pixels correspond to their distance to the building boundaries. Going even further, Zuo *et al.* [46] propose a *hierarchically fused FCN (HF-FCN)* which approaches a similar strategy as Yuan [45] by hierarchically fusing the information from multi-scale receptive fields of the network built on the basis of VGG-16 architecture. Maggiori *et al.* [47] convert the fully connected network proposed by Mnih [37] to FCN and generate a building mask out of RGB satellite imagery by, firstly, training the network on possibly inaccurate *OpenStreetMap (OSM)* data, and, finally, refining the model on a small amount of hand-labeled data. The major differences from our work are that the network architecture is much shallower and does not produce the output map of the same size as the input image. On the other hand, like in our work, the approach combines coarse and fine information from different layers in order to produce more detailed results. In continuation of their previous work, Maggiori *et al.* [48] investigate the network built on the basis of FCN proposed by Long *et al.* [13] combined with a *multi-layer perceptron (MLP)* on top of it. However, MLP is a FC network applied to every pixel individually and it significantly enlarges the number of parameters in the network. In our earlier work [49], we propose to use a deep learning network FCN8s developed by Long *et al.* [13] for segmenting the buildings only from nDSM data. In contrast to the current work, we copy the nDSM three times and initialize the network with the model pre-trained on RGB images. However, there is no influence on the final result, as the elevation information has a different statistics in comparison to spectral information and, thus, requires different feature representation.

An important milestone for semantic segmentation of remote sensing images with deep learning was multi-stream architectures that learn separate convolution layers for different data modalities. A study of Lagrange *et al.* [50] shows that combining image and DSM is essential for retrieving some specific classes. A further development of deeper networks and the late fusion of the spectral and height information were investigated in the work of Marmanis *et al.* [51]. This work is most closely related to ours, motivated by the interaction of multi-source information and integration of more detailed information from earlier layers to top. The difference from our work is an ensemble learning of the developed model which is a naive averaging after training the model with different initializations. This strategy is not integrated in our current work as we want to demonstrate the model strength to make good predictions after only one complete training. Besides, we do not engage the gradual training which does not guarantee the improvements of final results. A similar architecture strategy is approached by Sherrah [52]. However, in contrast to many deep learning architectures, this work presents a novel no-downsampling network to maintain the full resolution of the imagery at every layer in the FCN. This is achieved by using the "atrous" algorithm [53] which removes the pooling layers that caused the down-sampling effect. In contrast to our work, the fusion is done much earlier in the fully-connected layers. However, the fusion at this point did not lead to significant improvement. Besides, in opposite to our strategy, the authors did not up-sample the resulted output image from the network but used bilinear interpolation afterward to achieve the same size as the input image. Both works [51, 52] advocate using pre-trained networks for the spectral channels, but train the network for height channel from scratch. Audebert *et al.* [54] investigate the hybrid encoder-decoder architecture from Badrinarayanan *et al.* [55] for dealing with diverse data sources by concatenating the intermediate feature maps of separately trained dual-stream architecture and feeding the merged results to a three-convolution layers network. Besides, they introduce multi-kernel convolutional layers in the decoder part to aggregate multi-scale information while up-sampling. Although, their fusion network is similar to ours the main difference is the additional combination of the output from the fusion network with average scores of the two independent branches. In our case, the fusion is supposed to correct errors within one fusion network without additional concatenations by giving more weight to the activations of the most suitable information among complementary sources. Moreover, the presented architecture, in contrast to ours, does not have any "skip" connections which allow the decoder to recover important details that are lost due to the down-sampling in the encoder. Another difference to our work is the data they used. The addition to spectral image in this work is a composite image consisting of DSM, nDSM, and NDVI information. As NDVI is a good indicator for vegetation, the authors believe that this kind of auxiliary information helps to improve vegetation detection. But, as the components of the index calculation (the *infrared (IR)* and *red (R)* channels) are already given to the network as input, the network is capable to distinguish the vegetation itself. Another reason not to take NDVI into account, at least for the building detection, is that we do not need a precise vegetation prediction but only buildings discrimination from above-ground objects.

In the context of urban scene understanding, not only the DSM can provide complementary knowledge. Recently, efforts have been made for joint edge detection and semantic classification. Marmanis *et al.* [56] present an end-to-end ensemble of CNNs for semantic segmentation with an explicit awareness of semantically meaningful class boundaries. The boundary detection significantly improves semantic segmentation results and the overall accuracy achieved more than $>90\%$ on the ISPRS Vaihingen benchmark. Hu *et al.* [57] investigate the fusion of spectrum information of hyperspectral image and the scattering mechanism of PolSAR data. They propose a novel architecture which fuses two separated streams in a balanced manner. Since spaceborne remote sensing videos are becoming essential resources for remote sensing applications, Mou *et al.* [58] propose to fuse multi-spectral images and space videos for spatiotemporal analysis, to achieve a fine-resolution spatial scene labeling map.

Currently, *generative adversarial networks (GANs)* are also investigated in the remote sensing domain. Isola *et al.* [59] attempt to generate a mapping function to convert a satellite photo into a map and vice versa. Marmanis *et al.* [60] propose to use the GAN for artificial *synthetic aperture radar (SAR)* images generation in order to increase the training dataset.

In this paper, we explore the potential of multi-source data fusion, within one FCN architecture, for fully auto-

mated end-to-end building footprint extraction from high-resolution remote sensing images. Our contributions are the following:

- We efficiently adapt the FCN8s architecture developed by Long *et al.* [13] from generic everyday images to satellite images and analyze it for three different data sources: RGB, nDSM, and PAN images.
- We augment the FCN8s with additional "skip" connection, which combines the predictions at an earlier stage with the later one, for improving the segmentation results. We name the network FCN4s and inspect the improvements on RGB, nDSM, and PAN images in comparison to FCN8s.
- Inspired by the possibility to fuse multi-source data within one deep convolutional framework, we propose a Fused-FCN4s architecture which employes a late fusion approach of three identical parallel FCN4s networks, carrying information from RGB, nDSM, and PAN images. To our knowledge, this is the first work which applies in a direct way a deep convolutional architecture on RGB, nDSM, and PAN satellite data for building footprint extraction.
- As generalization is a key point for remote sensing applications, we demonstrate the generalization capability of the proposed network by applying it to a different urban landscape, unseen by the model before.

3 METHODOLOGY

3.1 Convolutional Neural Networks

CNNs are a category of artificial neural networks that have successfully been applied to visual imagery understanding. They are commonly organized in a series of layers. This hierarchy allows the network to learn multiple levels of data representation, starting from low-level features at the bottom layers, such as edges and corners, proceeding to generate coarse feature maps with high-level semantic information at the top layers. CNNs take advantage of the 2D structure of an input image by applying on it learnable 2D convolutional filters

$$y_j^l = \sigma\Big(\sum_{k\in -\frac{W}{2}\times\frac{W}{2}} w_{jk}\cdot y_k^{l-1} + b_j^l\Big) \tag{1}$$

which connect each neuron at level l with a specially localized region of fixed size $W \times W$ from previous layer $l-1$, and takes a weighted sum over all neurons followed by some activation function σ. The b_j^l corresponds to a bias. Due to the weights w_{jk} being shared across all neurons for each dimension per layer, the number of free parameters is significantly reduced in the model, compared to the standard MLP, which differs mainly by the fact that no weight sharing takes place in this type of neural networks. Additionally, the weight sharing introduces translation equivariance [61], another desirable attribute for the network. The bias can be considered yet another weight (with $y_{i=0} = 1$). The merit of the activation function is to introduce non-linearity into the network. The most common activation function applied after each convolutional layer in CNNs is the *rectified linear unit (ReLU)*

$$y_{relu}^l = \max(0, y^l) \tag{2}$$

which sets all negative numbers in the convolution matrix to zero and keeps the positive values unchanged

The main advantages of using ReLU in neural networks are, first, it induces sparsity in the hidden units, second, it does not suffer from the gradient vanishing problem [62].

As CNNs were originally developed for image classification problems, their goal was to predict the correct class associated with the input image. Therefore, the top layers of the network are usually FC layers, which merge the information of the whole image. The final layer is a 1D array and consists then of as many output neurons as there are possible classes, representing class assignment as probabilities, most often using $\mathrm{softmax}$ normalization on each of the neurons.

The classifier computed by the network is determined by the weights and biases parameters. To generate an optimal network classifier means to find such weights and biases which will minimize the difference between predicted values and target values. The misclassifications are penalized by a loss function $\mathcal{L}(\mathbf{x}, \mathbf{t}, \mathbf{p})$. The commonly used cross-entropy loss function

$$\mathcal{L}(\mathbf{x}, \mathbf{t}, \mathbf{p}) = -\sum_i t_i \log p(x_i) \tag{3}$$

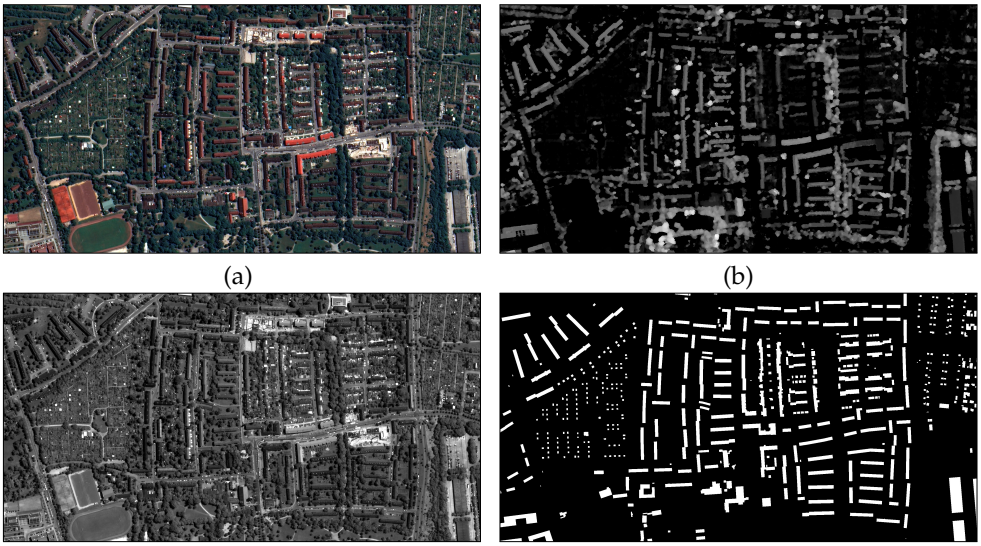
avoids the problem of slowing down the learning (in comparison to, for instance, the Euclidean distance loss function) and provides a more numerically stable gradient when paired with softmax normalization [47]. Here, $\mathbf{x} = \{x_1, \ldots, x_n\}$ is the set of input examples in the training dataset and $\mathbf{t} = \{t_1, \ldots, t_n\}$ is the corresponding set of target values for those input examples. The $p(x_i)$ represents the output of the neural network for given input x_i. We minimize the logistic loss of the $\mathrm{softmax}$ outputs over the whole patch.

A standard technique to minimize the loss function is *gradient descent* which computes the derivatives of the loss function with respect to parameters $\frac{\partial \mathcal{L}}{\partial w_i}$ and $\frac{\partial \mathcal{L}}{\partial b_i}$ and updates the parameters with learning rate λ in the following way:

$$w_i \leftarrow w_i - \lambda \frac{\partial \mathcal{L}}{\partial w_i} \tag{4}$$

$$b_i \leftarrow b_i - \lambda \frac{\partial \mathcal{L}}{\partial b_i} \tag{5}$$

The derivatives $\frac{\partial \mathcal{L}}{\partial w_i}$ and $\frac{\partial \mathcal{L}}{\partial b_i}$ are calculated by the *back-propagation* algorithm [63] commonly used in the *stochastic gradient descent (SGD)* optimization algorithm in small batches for efficiency. In this model, we used SGD with momentum, an extent to the vanilla SGD method. Additional methods have been suggested recently like *ADAM* [64] and *RMSProp* [65]. Although the optimization technique is very critical in the case of training from scratch, its role is muted in the case of pre-training, because the network is hindered from rapidly changing the weights, typically by using a very small learning rate. Therefore the technique itself plays finally a less important role in the convergence. A good



(a) (b) (c) (d)

Fig. 3: Test area: (a) RGB, (b) nDSM, (c) PAN and (d) Ground truth building mask.

overview of gradient descent optimization algorithms is given by Ruder [66].

3.2 Fully Convolutional Network Architecture

In this paper, we address a full pixel-wise binary labeling problem for building vs. non-building classes. It means that we want to give the network an image and receive an output image of the same size, with meaningful shape and structure of building footprints. The original CNNs were constructed for recognition tasks where only one label is assigned to each image. The recently developed FCNs became the state-of-the-art methodology for semantic segmentation. They are the extensions of the traditional CNN architecture, where all FC layers are replaced with convolutional layers. The advantage of this transformation is the independence of the input image size. Additionally, in contrast to the basic CNNs, FCNs do not lose the spatial information in the top layers but allow to track it back. The per-class probability maps $cl_i(x, y)$, which the FCNs generate, have a coarse resolution due to the pooling and convolution with stride larger than 1 operations along the network. The number of probability maps $cl_i(x, y)$ in the last convolutional layer is equal to the number of classes of the task. So, for our binary classification problem, this number is equal to 2. In order to up-sample the feature maps from the previous layer, the FCNs are augmented with "deconvolution" layers. This type of layer performs a learned interpolation from a set of nearby points. The construction of the network with several *deconvolution* layers at its top part allows obtaining the resulted class probability maps of the same size as the input image. In our network, we initialize the deconvolution weights with a set of bilinear interpolation parameters.

3.2.1 FCN4s Network

Applying several up-sampling layers and, as a result, bringing the classification maps to the original size, does not guarantee very detailed and accurate object boundaries in the resulting images. Long *et al.* [13] were the first who suggested to use the high-frequency information from the feature representations of the shallow part of the network, bypassing several layers of nonlinear processing, and combining it using an *element-wise addition* with the output from the deconvolution layers at the same resolution. This type of structure received the name of "skip" connection and is depicted in Figure 1 by a long arrow in violet color. In this way, the FCN8s network proposed by Long *et al.* [13] hierarchically includes the earlier layers pool4 and pool3 to the upper layers of the network, adding more detailed information.

However, the FCN8s was originally created for semantic labeling in the field of computer vision, where objects are big and well separated. Remote sensing imagery, in contrast to multimedia images, is very different. First of all, due to the big difference in the *ground sampling distance (GSD)*, even if the resolution of remote sensing images is high, still, the containing information is very heterogeneous. It consists of many objects like trees, buildings, roads, etc. Secondly, those objects can be represented only by a small number of pixels. Therefore, it is more challenging to extract very accurate boundaries and structures from such images. As a result, we modify the FCN8s network to an FCN4s by adding yet another "skip" connection from pool2 layer, which incorporates even finer details, allowing more efficient building footprint reconstruction (see Figure 1). We also adapt the number of channel dimensions from 21 to 2. The training is done by fine-tuning the weights of the model, which is pre-trained on the large image collection of ImageNet.

3.2.2 Fused-FCN4s Network

For the semantic segmentation task, the data used are often three-channel imagery. In this work, we propose a new network which integrates image information from RGB and PAN images, together with depth information from nDSM, as the latter provides geometrical silhouettes, which allow a better separation of buildings from the background.

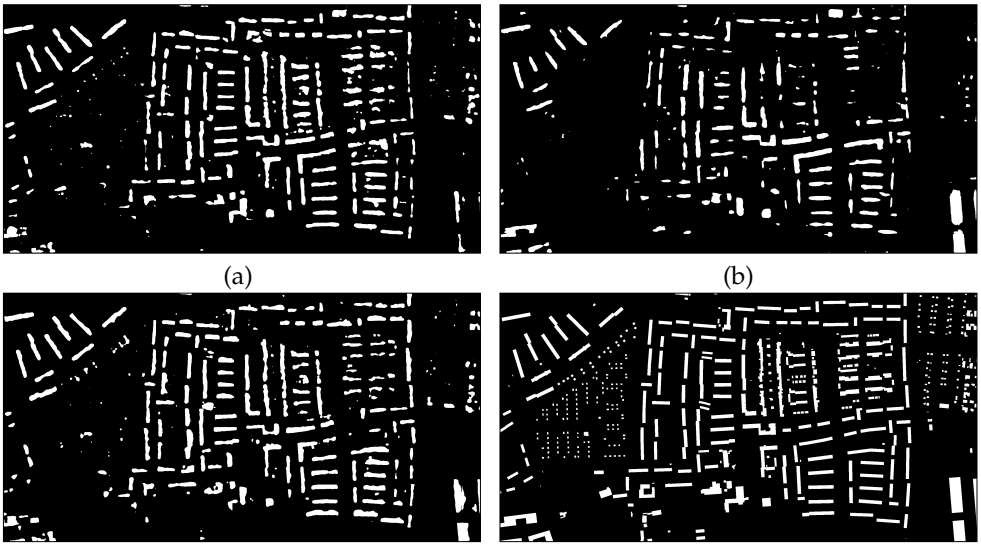


Fig. 4: The relative performance of the FCN8s model for building mask generation on individual data sources: (a) RGB, (b) nDSM and (c) PAN images. Image (d) illustrates ground truth.

Besides, depth images are invariant to illumination and color variations. Since depth information and intensity have different physical meaning, we propose a hybrid network where three separate networks with the same architecture are used: We feed one part with the red, green, and blue spectral bands and initialize it with the weights pre-trained on ImageNet as mentioned in Section 3.2.1. The second part we feed with the PAN image converted to three-channel by copying it three times. The network is initialized the same way as the first part. The reason to use pre-trained weights for gray scale image is twofold: First, the pre-trained networks demonstrate a strong ability to generalize to images outside the Imagenet dataset via transfer learning. Thus, we make modifications in the pre-existing model by fine-tuning it. Second, the PAN image has the same topology as our RGB image. So, as the visual filters from generic images can be built upon for RGB images, they are applicable for PAN images too. The third branch is fed with one-channel nDSM, initializing the convolutional layers randomly since elevation data and intensity data have different modalities and, as a result, require different feature representations. We examine two fusion strategies: a) a naïve averaging of three branches after softmax, and b) merging by the neural network itself.

The schematic diagram of the proposed network architecture is illustrated in Figure 2. First, it stacks the sets of spectral and height features from three streams at a very top level, but before the last *up_4* up-sampling layer as depicted in Figure 2. As a result, the number of features increases three times. Second, the up-sampling is applied to bring the combined feature maps to the final size. Finally, the resulting intermediate features are sent as an input to three additional convolutional layers of size 1×1, which play the role of information fusion from different modalities, and can correct small deficiencies in the predictions, by automatically learning which stream of the network gives the best prediction result. This architecture is similar to the one presented by Marmanis *et al.* [51]. Although our implementation is based on the paper description, we made additional modifications which experimentally improve our final results. For example, the number of feature maps at the higher layers of the network is set to a larger number to allow the network to learn a wider range of features. However, we decreased the number of channels suggested by Marmanis *et al.* [51] from 60 to 30 and, experimentally, obtained better results. Besides, having a network with a huge number of parameters but rather small training set can lead to overfitting. Additionally, in contrast to Marmanis *et al.* [51], we did not find it necessary to introduce *local response normalization (LRN)* to the last layer of three independent branches for spectral intensities and height before merging as the network is able itself to balance the activations between heterogeneous data. It also prevents from additional tuning of the hyper-parameters for LRN.

The network can only see a part of the image when it is centered at a pixel. This region in the input is the receptive field for that pixel and can be computed by the formula mentioned in Le *et al.* [67]

$$R_k = R_{k-1} + (f_k - 1) \prod_{i=1}^{k-1} s_i, \tag{6}$$

where R_k is the current layer, R_{k-1} is the previous layer, f_k represents the filter size of layer k, s_i is the stride of layer i. The receptive field of the output unit of the network that we use in this work is 404×404 pixels.

4 Study Area and Experiments

We performed experiments on WorldView-2 data showing the Munich city, Germany, consisting of a color image with red, green, and blue channels, a very high-resolution stereo PAN imagery and a DSM derived from it using the *semi-global matching (SGM)* method [68]. The RGB and PAN images used in the experiment have been ortho-rectified,

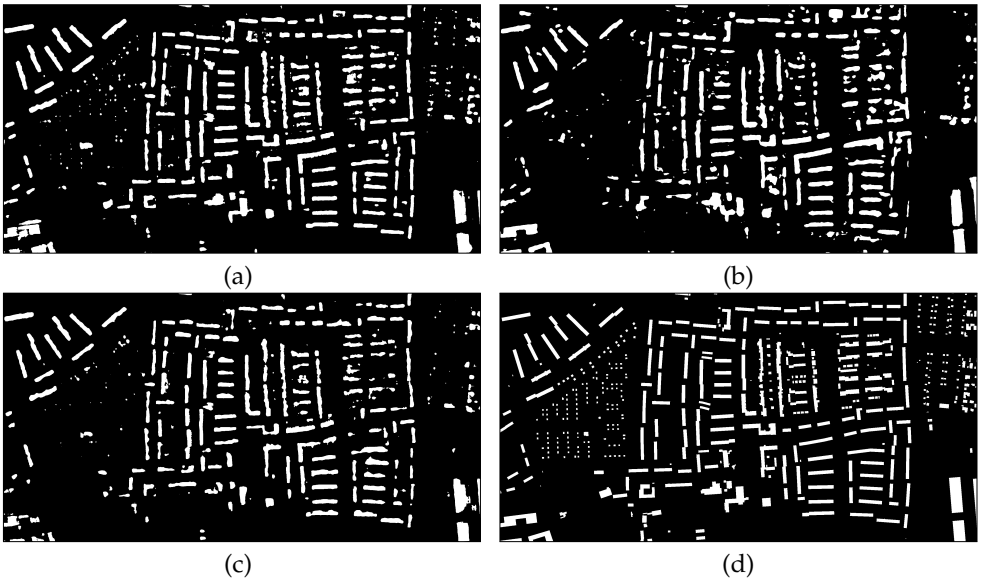


Fig. 5: The relative performance of the FCN4s model for building mask generation on individual data sources: (a) RGB, (b) nDSM and (c) PAN images. Image (d) illustrates ground truth.

because it is important for building detection to have images where every pixel in the image is depicted as if viewed at nadir, so that occlusions do not pose a challenge.

As a ground truth for our training, a building mask from the municipality of the city of Munich, covering the same region as the satellite imagery, is used for learning the parameters in the neural network.

In order to investigate the prediction model capacity over a different urban landscape, a second WorldView-2 dataset showing a small part of Istanbul city, Turkey, was considered. As the ground truth for this area is not available, a building mask was extracted from OSM. However, only a few building footprints are available for this area and the rest is missing. Therefore, a small area of around 0.5 km^2 was selected over the available building footprints. The rest was manually delineated.

4.1 Data Pre-processing

To perform a network training from the multiple data sources, first a PAN image with a GSD of 0.5 m was used to pansharpen the color image with a GSD of 2 m using the pansharpening method proposed by Krauß *et al.* [69].

Secondly, in order to obtain above-ground information only, namely to generate a nDSM, the topographical information was removed from DSM based on the methodology described by Qin *et al.* [70]. Additionally, by investigating the histogram of height data in the nDSM, it was found that there are about 0.05% outliers, which enlarge the distribution range dramatically (to 205 m height), although the majority of values lay within a much smaller range. The explanation to these outliers can be the presence of noise, due to the absence of information because of clouds. Therefore, the decision was made to remove this 0.05% of outliers and use linear spline interpolation to find the values of thresholded points. It should be mentioned that even if there are some buildings in the image higher than the selected threshold, for our binary classification task it is not very critical to loose the true height of very high buildings within the city area, since we are only interested in footprints. Another advantage of the suggested data pre-processing is the simplicity of the network training.

4.2 Implementation and Training Details

We developed our FCN4s and Fused-FCN4s models based on the FCN8 implemented in *Caffe* deep learning framework [71]. For learning process, we prepared the training data consisting of 22057 pairs of patches, and validation data of 3358 pairs, selected from a different area. The patches cropped from the satellite image have a size of 300×300 pixels. Having a large receptive field size of the architecture leads to the question about the relative influence of boundary effects on the predictions. In our case, as the context information is available only within 300×300 pixels, each output unit of the network is influenced by the boundary effect. Therefore, to prevent artifacts and discontinuity at patch boundaries, we used an overlap of 200 pixels out of 300 (67 %) when sliding the window across the satellite image in both directions. To further improve the prediction on boundaries, all overlapping patches are stacked together first, then the final prediction is calculated as the average at each pixel. As a result, some pixels are predicted once, twice or four times like the ones at the corners. This is a commonly used approach for remote sensing problems [54].

As mentioned in Section 3.2.2, the two branches of the network corresponding to spectral images were initialized with a pre-trained model. This applies to the network before the fully convolutional layers. All layers above the fully convolutional layers were initialized within a range defined inversely proportional to the number of input neurons. For a layer with N neurons, the weights were initialized in the range $[-\frac{1}{N}, \frac{1}{N}]$ using uniform sampling. The network branch corresponding to nDSM data is trained from scratch for the reasons explained in Section 3.2.2. We start the training process of our network with learning rate $\lambda = 0.01$

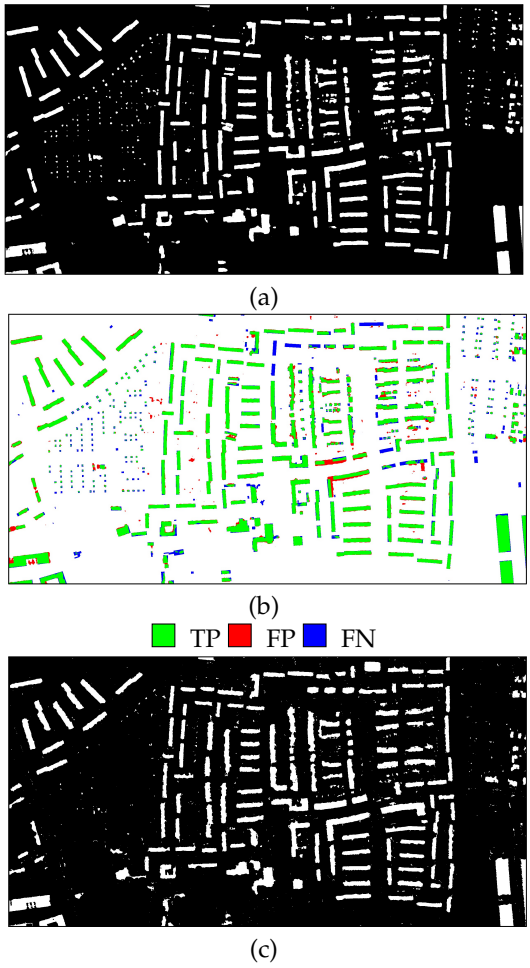


Fig. 6: The comparison of generated building mask over test area obtained (a) directly form Fused-FCN4s and (c) from Krauß *et al.* [69]. Image (b) depicts the extracted building footprints in respect to reference building footprints of Fused-FCN4s.

for all randomly initialized layers and λ = 0.001 for layers initialized with the pre-trained model, decreasing them by a factor of 10 for each 20000 iterations. The total number of iterations was set to 60000 with batch size of 1 on a single NVIDIA TITAN X (Pascal) GPU with 12 GB memory. A weight decay η and momentum factor m were set to η = 0.0005 and m = 0.9, respectively. All parameters were obtained empirically during investigation of the training process on the validation dataset. Within the training, random shuffling of the samples was performed before feeding them into the network.

4.3 Comparison with alternative methods

Apart from the developed FCN4s network, presented in Section 3.2.1, we also compare our approach with the FCN8s network proposed in [13]. We directly employed it for RGB and nDSM images, by changing only the number of outputs to 2 in order to be consistent with our binary classification task. During the fine-tuning of the FCN8s on RGB and PAN images using the pre-trained Imagenet model, the base learning rate was set to λ = 0.0001. For training the FCN8s from scratch for nDSM image, the base learning rate was set to λ = 0.01.

In order to demonstrate the advantage of end-to-end deep learning data fusion, we compare the designed architecture with naïve prediction fusion. Moreover, to indicate the influence of every data source we compare our approach with two-stream fusions: 1) RGB and nDSM; 2) PAN and nDSM.

Besides, we conduct a comparison on DSM-based building detection method proposed by Krauß *et al.* [69]. This method, first, generates a height map by distinguishing the above ground objects from the ground level ones using nDSM. The extracted height map is used then for buildings delineation from the surroundings by applying the Advanced Rule-based Fuzzy Spectral Classification [69]. The implementation distributed by the authors is applied to the nDSM and 8-channel multi-spectral image covering the test area.

5 Results and Discussion

In the following section, the results of the considered experiments for FCN8s, FCN4s, and the proposed Fused-FCN4s, on different data sources, is presented. Their respective performance is discussed, in order to evaluate the introduced architecture for binary building mask generation, both qualitatively and quantitatively. To demonstrate the effectiveness of the models, we fed a new test dataset to the network, unseen before neither for training nor for the validation. A test area from the city of Munich and its corresponding ground truth image is depicted in Figure 3.

5.1 Qualitative Evaluation

5.1.1 FCN8s Network

The building masks generated by the FCN8s network separately on RGB, nDSM, and PAN images are presented in Figure 4. As can be seen from the results, the FCN8s model, generated for multimedia imagery semantic segmentation, is applicable to remote sensing data too. Moreover, not only intensity images but also the nDSM representing depth information can be used for building footprints extraction using FCNs. This has been also analyzed by Davydova *et al.* [38] and Bittner *et al.* [49]. As illustrated in the figures, the FCN8s model is able to extract the buildings from each given data source without any influence of other above-ground objects such as trees, cranes etc. However, as it can be noticed, some footprints are better extracted from intensity images and some of them from the depth image. For example, there are two big buildings in the bottom right corner. Referring to the original RGB image in Figure 3a, one can see that the roofs of both constructions have a color similar to the asphalt. Therefore, we deduct that the network confuses these buildings with the road. From PAN images, the network could learn different features and, as a result, enable the network to identify the area as buildings, but not optimally yet. On the other hand, from the height information provided by the nDSM, it was easier for the network to distinguish these buildings from the ground. As can be seen from the results, many buildings are missing in

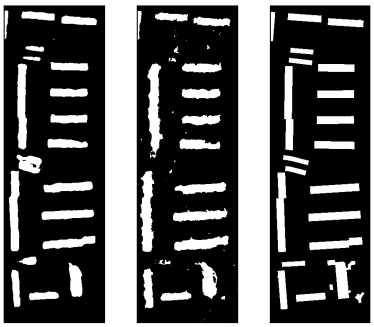
	Mean acc.	Mean IoU	Overall acc.	IoU	$F_{measure}$
FCN8s					
RGB	82.8	75.5	94	57.6	73.1
nDSM	74.5	69	92.8	45.7	62.7
PAN	84.6	77	94.3	60.2	75.1
FCN4s					
RGB	89.3	81	95.2	67.1	80.4
nDSM	83.3	73.3	92.9	54.3	70.4
PAN	84.4	77.5	94.6	60.9	75.7
Fused-FCN4s					
RGB & nDSM	90.9	84.7	96.1	73.5	84.7
PAN & nDSM	87.5	82.2	95.9	68.9	81.6
RGB & nDSM & PAN	**91.5**	**86**	**96.8**	**75.7**	**86.1**
Naïve fusion					
RGB & nDSM & PAN	87.6	81.7	95.7	68.1	81
DSM-based building detection method					
MS image & nDSM	89.1	78.2	94.6	62.4	76.8

TABLE 1: The quantitative evaluation of proposed Fused-FCN4s on three data sources in comparison to different methodologies and setups.

the building mask, even the one extracted from the nDSM. This can be caused by trees occluding some buildings, or inaccurate height data in these locations.

5.1.2 FCN4s Network

It is always good to have additional information which can be added to the system, as it makes the system more powerful. CNNs are capable of extracting representative features for a classification task if enough information is present. Therefore, as we wanted to improve the building outlines without any post-processing steps, it was decided to enrich the system by adding more detailed information from earlier network layers. As a general rule, CNNs gradually abandon lower level features in the pursuit of higher levels, which leads to a more abstract description of the image. This strategy can be countered by passing lower level features up the hierarchy in a separate path (skip connection). In this way, the network itself automatically learns higher detailed building representations. The effectiveness of the suggested FCN4s approach is illustrated in Figure 5. First of all, in each resulting image, for every data source, one can notice that more buildings are extracted. Second, the shapes of the footprints, even for the complex building structures, are closer to the ground truth and better in comparison to the one extracted by FCN8s architecture. Finally, the addition of the pool2 skip connection, enable the network to recognize even the low-rise buildings.



(a) (b) (c)

Fig. 7: The detailed comparison between (a) Fused-FCN4s and (b) DSM-based building detection method proposed by Krauß *et al.* [69]. Image (c) depicts ground truth.

5.1.3 Fused-FCN4s Setup

In this section, we investigate different setups of Fused-FCN4s architecture. Setting the number of convolution layers to 2 for performing a fusion from different network streams and increasing the number of feature maps at the top layers lead to a tendency to improve the result (see Table 2). This happens due to the fact that the increase of the parameters number in the network raises its capacity and, thus, makes it possible to perform better generalization. However, at some point the network can reach too much complexity which comes with the risk of overfitting. This effect can be observed with a configuration of 60 feature maps and three convolutional layers. The results of generalization degrade in comparison to a fusion network with 30 feature maps and three convolutional layers. Growing the number of feature maps in the network increases the computation time respectively as depicted in Table 2. However, it helps to improve the results significantly. Hence, we choose the model with 30 feature maps and 3 convolutional layers as it provides the best results in this experiment.

5.1.4 Fused-FCN4s vs. FCN8s and FCN4s

The Fused-FCN4s architecture, which combines the spectral information from RGB and PAN images, together with the height information from nDSM, delivers the best performance in discriminating buildings from background, in comparison to FCN8s and FCN4s shown in Figures 4 and 5, respectively. The results obtained by Fused-FCN4s architecture are shown in Figure 6. For visualization and better interpretation the extracted building footprints are also overlapped with the reference building footprints in Figure 6b. The significant improvement of the buildings outlines can be easily observed. The footprints are more accurate and their shapes are more complete without missing parts of the various structures. It also can be seen that the network really benefits from all data sources, which allow it to extract more detailed information of building construction compared to the reference image in Figure 3d. For example, the building in the left bottom corner obviously has some additional

	Mean acc.	Mean IoU	Overall acc.	IoU	$F_{measure}$	n_p	t_f, ms	t_b, ms	t_{f-b}, ms
2fmaps_2conv	81.4	74.5	93.9	69.7	71.6	402 773 872	85.86	347.648	433.647
30fmaps_2conv	91	85	96.5	74	85	403 205 352	93.1116	367.919	461.177
60fmaps_2conv	91.4	85.9	96.7	75.4	86	403 672 872	102.791	376.647	479.616
2fmaps_3conv	90.7	84.6	96.3	72.83	84.3	402 773 876	86.41	350.601	437.16
30fmaps_3conv	**91.5**	**86**	**96.8**	**75.7**	**86.1**	403 205 772	93.5415	370.621	464.297
60fmaps_3conv	90.9	85.6	96.7	74.9	85.7	403 647 612	102.826	377.569	480.529

TABLE 2: The results of detailed investigation on Fused-FCN4s model performance with respect to modifications in architecture. We vary the number of feature maps (fmaps) in the top layers together with the number of convolutional layers after merging the streams from three data sources. The n_p indicates a number of parameters in the network, t_f is the average time for one forward pass on a single NVIDIA Titan X (Pascal) GPU, t_b is the average time for one backward pass and t_{f-b} is the average time for one forward-backward pass.

structures in the middle, which can be easily identified on the nDSM image, but they are missing in the ground truth. The extraction of low-rise buildings, on which the selected scene is rich, is more accurate now, and their pattern of placement is very close to the ground truth. Some of them are still missing, but that is explainable due to their really small size, difficult to distinguish even for the human eye.

Besides, it is experimentally proven that the proposed network benefits from three remote sensing images used for training in comparison to two-stream networks of RGB and nDSM and PAN and nDSM (see Table 1). We can see that the use of the PAN image leads to improvements of 2.2% on *intersection over union (IoU)* and from 0.7% to 2% on the rest of the metrics.

5.1.5 Fused-FCN4s vs. Naïve Fusion

The experimental results from Table 1 demonstrate that naïve fusion by averaging the predicted maps improves the IoU metrics only by 1% in comparison to the results, achieved by FCN4s model trained on RGB. But the proposed Fused-FCN4s boosts the IoU metrics by 8%. Thus, the shapes of generated building footprints are enhanced in comparison to those obtained by single FCN4s. Additionally, a significant improvement of other metrics is also achieved. This proves that the network learns by itself from which multi-source data the better prediction of the pixel can be gained.

5.1.6 Fused-FCN4s vs. DSM-based building detection method

As can be seen from Figure 6c the DSM-based building detection method proposed by Krauß *et al.* [69] is able to extract a similar building mask as our proposed approach. However, a close investigation shows that Fused-FCN4s is able to find more buildings than the DSM-based building detection method (see Figure 7). Additionally, one can notice that the extraction of low-rise buildings by our approach is significantly better. Besides, the footprints outlines are more accurate and rectilinear, that makes them look qualitatively more realistic and, as a result, similar to the ground truth.

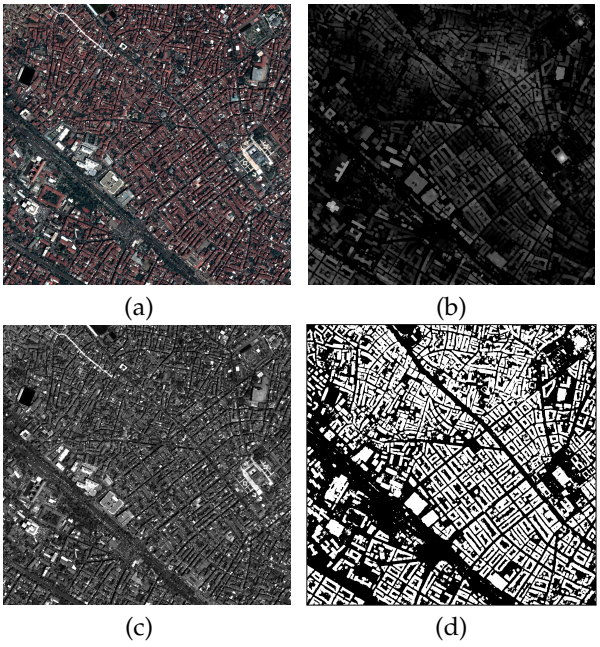


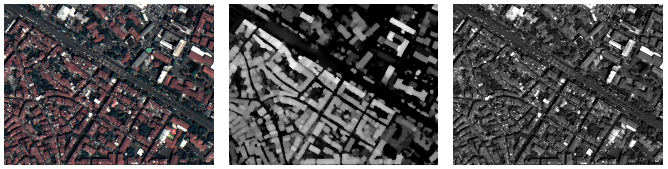
Fig. 8: Prediction over Istanbul city, Turkey on WorldView-2 data. (a) RGB image; (b) nDSM; (c) PAN. (d) The resulted mask from Fused-FCN4s.

5.2 Quantitative Evaluation

For quantitative evaluation of the obtained results, we evaluated the metrics commonly used in semantic segmentation problem. The first group of metrics is described in Long *et al.* [13]. They are *mean accuracy*, *mean IoU* and *overall accuracy*

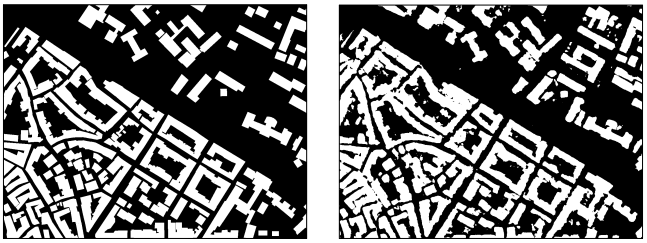
FCN4s					
	Mean acc.	Mean IoU	Overall acc.	IoU	$F_{measure}$
RGB	84.3	72.8	85.1	66.9	80
nDSM	76.3	60	75.2	54	70.7
PAN	79.4	66.6	81.3	58.8	74.1
Fused-FCN4s					
RGB & nDSM	85	72.8	84.9	67.7	80.8
PAN & nDSM	84.9	72.6	84.8	66.6	79.3
RGB & nDSM & PAN	**85.1**	**73.5**	**85.5**	**68.1**	**81**

TABLE 3: Prediction accuracies of FCN4s and Fused-FCN4s models on all investigated metrics over Istanbul dataset



(a) (b) (c)

Fig. 9: The selected area over Istanbul city for statistical evaluation. (a) RGB; (b) nDSM; (c) PAN



(a) (b)

Fig. 10: Small area of initial Istanbul dataset. Image (a) shows the ground truth, partially obtained from OSM and partially completed by manually drawing the footprints. Image (b) illustrates the predicted map.

$$\text{Mean accuracy} = \frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i}, \tag{7}$$

$$\text{Mean IoU} = \frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}, \tag{8}$$

$$\text{Overall accuracy} = \frac{\sum_i n_{ii}}{\sum_i t_i}, \tag{9}$$

where n_{ij} is the number of pixels belong to class i, but predicted as class j, n_{cl} is the number of different classes, and $t_i = \sum_j n_{ij}$ is the total number of pixels belong to class i.

The second group of selected metrics, suitable for binary classification evaluation, are based on predicted values represented by the total number of true positive (TP), false positive (FP) and false negative (FN). Based on these values the *F-measure* is defined as

$$F_{measure} = \frac{(1+\beta^2)TP}{(1+\beta)^2 TP + \beta^2 FN + FP}, \tag{10}$$

where for our work the parameter β was set to 1. Additionally, we use the IoU metric

$$\text{IoU} = \frac{TP}{TP + FN + FP} \tag{11}$$

adapted for the task, where the amount of pixels belonging to the objects (buildings) are much smaller compared to those belonging to the background. This metric is represented by the proportion of the number of pixels classified as buildings, both in the predicted image and in the ground truth, to the total number of pixels classified as buildings in each of them [47].

The summarized performances of FCN8s, FCN4s, Fused-FCN4s networks and DSM-based building detection method proposed by Krauß *et al.* [69] using above described metrics are grouped in Table 1. From the quantitative statistics we can see that, first, the performance of all networks on spectral images are better than on the image representing the height information. This is reasonable, as the DSM images themselves are obtained from the multi-view stereo PAN pair and some information can be unavailable, due to occlusions by different objects or clouds within the scene. Second, by further augmenting the architecture with "skip" connection from the pool2 layer, to generate FCN4s network, we gain improvements of performance on nDSM and RGB images. However, for PAN image the improvement is not very significant. This is due to the fact that the network became more complicated using the additional connection as a result of an enlarged number of parameters, but the extracted information comes only from the three times duplicated image and is not enough to provide the network with much more features. Finally, the proposed Fused-FCN4s network obtains the best performance for all metrics in comparison to other networks and the DSM-based building detection method. The overall accuracy gained 2% points in comparison to FCN8s for RGB and PAN images, and around 4 % points related to FCN8s for the nDSM image. It should be mentioned, that the IoU metric on Fused-FCN4s network increased over 15% and 30% in comparison with FCN8s on spectral and depth images, respectively. That indicates a significant improvement of the building footprint delineation accuracy. Besides, the difference of the IoU metric of 13.3% between the Fused-FCN4s and DSM-based building detection method, in favor of the first, points out that applying our approach there is no need for any post-processing steps for building outline refinement as it already provides very accurate building mask.

Processing a selected test area of 1300×2500 pixels with Fused-FCN4s network takes 25.89 seconds on a single NVIDIA Titan X Pascal GPU with a 100 pixels stride and around 2 minutes for stitching the overlapped patches for the final full image generation.

5.3 Model Generalization Capability

In order to investigate the model capacity to capture the essential features separating buildings from non-buildings, Istanbul dataset was used (see Figure 8(a)-(c)). This dataset is very different from the Munich dataset, and it is very challenging in itself due to the dense placement of buildings, and the vastly different construction and architecture style. Without re-training the model on the new dataset, the building footprint map was directly obtained by passing the WorldView-2 data through the FCN4s and Fused-FCN4s networks. From the resulting mask shown in Figure 8(d), it can be seen that the proposed model managed to predict reasonable building mask even from a new and quite complicated dataset. As it was mentioned in Section 4, for quantitative evaluation a small area of around 0.5 km^2 was selected (see Figure 9). The predicted results and ground truth of this area are presented in Figure 10.

The statistical results of the experiment over the small area can be found in Table 3. We can see that the model

achieves high performance on this dataset as well. Besides, the advantage of using fused data vs. only one is also demonstrated in Table 3.

It can be clearly seen that the model successfully extracts the shapes of building footprints, without missing any of them. The IoU metric confirms this statement by its high value of about $\sim 68.1\%$. Additionally, no influence of other above-ground objects such as trees is observed. However, one can notice a small improvement between using one spectral image or two together with an nDSM. Both RGB & nDSM and PAN & nDSM models already gave good results using the advantages from spectral and height information. Inserting additional spectral information only helps to improve minor errors, especially on building outline as the IoU shows high values. But it is still a significant progress as commonly used methodologies for building extraction are not very flexible and can not be easily generalized on different city areas. Moreover, it can be identified that the quantitative results are lower than the ones from Munich dataset. This can be explained by scene complexity: The network did not experience such types of constructions, their close placement to each other and the narrow streets. Besides, the maximum height within Munich nDSM area is 58.37 m and for Istanbul is 24.66 m which also can influence the performance. Another reason is that the manually generated ground truth is far from ideal, due to the subjective interpretation of human. The probable solution to those small problems can be a fine-tuning of the proposed model on some small areas of different cities, which will contribute to the model performance by introducing a new dataset for model learning, even if it is only a small part of the area.

6 Conclusion

We presented a novel method to segment buildings in complex urban areas using multiple remote sensing data on the basis of fully convolutional networks. The designed end-to-end Fused-FCN4s framework integrates the automatically learned relevant contextual features from spectral and height information from RGB, nDSMs, and PAN images respectively, within one architecture for pixel-wise classification, and produces a unique binary building mask. Both, spectral images and nDSMs, have their strong and weak sides, but they can complement each other significantly, as, for example, the nDSMs provides elevation information of the objects, but spectral images provide texture information and more accurate boundaries. The trained system was tested on two unseen areas of Munich city, Germany, and Istanbul city, Turkey, and achieved accurate results. Experimental results have shown that even small objects with tiny details in their building footprint can be successfully extracted from satellite images by applying the deep neural network framework. The proposed architecture can be generalized over diverse urban and industrial building shapes, without any difficulties due to their complexity and orientation. Additionally, we show that the designed model does not need any post-processing. Some noise or still present inaccuracies in the resulting building mask can be a result of buildings totally covered by trees, or very complex areas which are difficult to recognize even for the human eye, for accurately extracting the building outlines. Besides, a noisy nDSMs can influence the results to a great extend, as the height information is crucial to identify buildings. We believe that the presented technique has a great potential to provide a robust solution to the problem of building footprint extraction from remote sensing imagery at a large scale.

Acknowledgments

The authors would like to thank Jiaojiao Tian for providing her software for normalized Digital Surface Model (nDSM) generation, Pablo d' Angelo for providing the WorldView-2 data for Munich and Istanbul cities and his continuous support on processing these images.

References

[1] S. Ahmadi, M. V. Zoej, H. Ebadi, H. A. Moghaddam, and A. Mohammadzadeh, "Automatic urban building boundary extraction from high resolution aerial images using an innovative model of active contours," *International Journal of Applied Earth Observation and Geoinformation*, vol. 12, no. 3, pp. 150–157, 2010.

[2] G Sohn and I. Dowman, "Extraction of buildings from high resolution satellite data," *Automated Extraction of Man-Made Objects from Aerial and Space Images (III). Balkema Publishers, Lisse*, pp. 345–355, 2001.

[3] C. Ünsalan and K. L. Boyer, "A system to detect houses and residential street networks in multispectral satellite images," *Computer Vision and Image Understanding*, vol. 98, no. 3, pp. 423–461, 2005.

[4] X. Huang and L. Zhang, "Morphological building/shadow index for building extraction from high-resolution imagery over urban areas," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 1, pp. 161–172, 2012.

[5] A. O. Ok, "Automated detection of buildings from single vhr multispectral images using shadow information and graph cuts," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 86, pp. 21–40, 2013.

[6] C. Lin and R. Nevatia, "Building detection and description from a single intensity image," *Computer vision and image understanding*, vol. 72, no. 2, pp. 101–121, 1998.

[7] Z. Kim and R. Nevatia, "Uncertain reasoning and learning for feature grouping," *Computer Vision and Image Understanding*, vol. 76, no. 3, pp. 278–288, 1999.

[8] M. Brédif, O. Tournaire, B. Vallet, and N. Champion, "Extracting polygonal building footprints from digital surface models: A fully-automatic global optimization framework," *ISPRS journal of photogrammetry and remote sensing*, vol. 77, pp. 57–65, 2013.

[9] G. Sohn and I. Dowman, "Data fusion of high-resolution satellite imagery and lidar data for automatic building extraction," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 62, no. 1, pp. 43–63, 2007.

[10] D. H. Lee, K. M. Lee, and S. U. Lee, "Fusion of lidar and imagery for reliable building extraction," *Photogrammetric Engineering & Remote Sensing*, vol. 74, no. 2, pp. 215–225, 2008.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[12] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[14] A. Huertas and R. Nevatia, "Detecting buildings in aerial images," *Computer Vision, Graphics, and Image Processing*, vol. 41, no. 2, pp. 131–152, 1988.

[15] R. B. Irvin and D. M. McKeown, "Methods for exploiting the relationship between buildings and their shadows in aerial imagery," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 19, no. 6, pp. 1564–1575, 1989.

[16] R. Guercke and M. Sester, "Building footprint simplification based on hough transform and least squares adjustment," in *Proceedings of the 14th workshop of the ICA commission on generalisation and multiple representation, Paris*, 2011.

[17] D. H. Douglas and T. K. Peucker, "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature," *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 10, no. 2, pp. 112–122, 1973.

[18] K. Karantzalos and N. Paragios, "Recognition-driven two-dimensional competing priors toward automatic and accurate building detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 1, pp. 133–144, 2009.

[19] B. Sirmacek, H. Taubenbock, P. Reinartz, and M. Ehlers, "Performance evaluation for 3-d city model generation of six different dsms from air-and spaceborne sensors," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 1, pp. 59–70, 2012.

[20] F. Dornaika, A. Moujahid, A. Bosaghzadeh, Y. El Merabet, and Y. Ruichek, "Object classification using hybrid holistic descriptors: Application to building detection in aerial orthophotos," *Polibits*, no. 51, pp. 11–17, 2015.

[21] H. Baluyan, B. Joshi, A. Al Hinai, and W. L. Woon, "Novel approach for rooftop detection using support vector machine," *ISRN Machine Vision*, vol. 2013, 2013.

[22] D. Chen, S. Shang, and C. Wu, "Shadow-based building detection and segmentation in high-resolution remote sensing image.," *journal of multimedia*, vol. 9, no. 1, 2014.

[23] T.-T. Ngo, C. Collet, and V. Mazet, "Automatic rectangular building detection from vhr aerial imagery using shadow and image segmentation," in *Image Processing (ICIP), 2015 IEEE International Conference on*, IEEE, 2015, pp. 1483–1487.

[24] M. Liggins II, D. Hall, and J. Llinas, *Handbook of multisensor data fusion: theory and practice*. CRC press, 2017.

[25] H. C. Lai, R. Yang, and G. W. Ng, "Enhanced self-organizing map for passive sonar tracking to improve situation awareness," in *Information Fusion, 2007 10th International Conference on*, IEEE, 2007, pp. 1–7.

[26] A. Birk, N. Vaskevicius, K. Pathak, S. Schwertfeger, J. Poppinga, and H. Buelow, "3-d perception and modeling," *IEEE robotics & automation magazine*, vol. 16, no. 4, 2009.

[27] L. Matthies, Y. Xiong, R Hogg, D. Zhu, A Rankin, B. Kennedy, M. Hebert, R Maclachlan, C. Won, T. Frost, *et al.*, "A portable, autonomous, urban reconnaissance robot," *Robotics and Autonomous Systems*, vol. 40, no. 2, pp. 163–172, 2002.

[28] C. Chen, R. Jafari, and N. Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition," *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 4405–4425, 2017.

[29] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[30] G. L. Foresti and C. S. Regazzoni, "Multisensor data fusion for autonomous vehicle navigation in risky environments," *IEEE Transactions on Vehicular Technology*, vol. 51, no. 5, pp. 1165–1185, 2002.

[31] T. N. N. Hossein, S. Mita, and H. Long, "Multisensor data fusion for autonomous vehicle navigation through adaptive particle filter," in *Intelligent Vehicles Symposium (IV), 2010 IEEE*, IEEE, 2010, pp. 752–759.

[32] P. Gamba, "Image and data fusion in remote sensing of urban areas: Status issues and research trends," *International Journal of Image and Data Fusion*, vol. 5, no. 1, pp. 2–12, 2014.

[33] F. Rottensteiner, J. Trinder, S. Clode, and K. Kubik, "Building detection by fusion of airborne laser scanner data and multi-spectral images: Performance evaluation and sensitivity analysis," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 62, no. 2, pp. 135–149, 2007.

[34] G. Shafer *et al.*, *A mathematical theory of evidence*. Princeton university press Princeton, 1976, vol. 1.

[35] S. Zabuawala, H. Nguyen, H. Wei, and J. Yadegar, "Fusion of lidar and aerial imagery for accurate building footprint extraction," *Image Processing. Machine Vision Applications II*, vol. 7251, 72510Z–1, 2009.

[36] A. Turlapaty, B. Gokaraju, Q. Du, N. H. Younan, and J. V. Aanstoos, "A hybrid approach for building extraction from spaceborne multi-angular optical imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 1, pp. 89–100, 2012.

[37] V. Mnih, "Machine learning for aerial image labeling," PhD thesis, University of Toronto (Canada), 2013.

[38] K. Davydova, S. Cui, and P. Reinartz, "Building footprint extraction from digital surface models using neural networks," in *Proceedings of SPIE*, vol. 10004, 2016, pp. 1–10.

[39] K. Fukushima and S. Miyake, "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition," in *Competition and cooperation in neural nets*, Springer, 1982, pp. 267–285.

[40] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[41] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Y. Ng, "Convolutional-recursive deep learning for 3d object classification," in *Advances in Neural Information Processing Systems*, 2012, pp. 656–664.

[42] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013.

[43] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2015, pp. 234–241.

[44] S. J. M. Drozdzal, D. Vazquez, and A. R. Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation,"

[45] J. Yuan, "Automatic building extraction in aerial scenes using convolutional networks," *arXiv preprint arXiv:1602.06564*, 2016.

[46] T. Zuo, J. Feng, and X. Chen, "Hf-fcn: Hierarchically fused fully convolutional network for robust building extraction," in *Asian Conference on Computer Vision*, Springer, 2016, pp. 291–302.

[47] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 2, pp. 645–657, 2017.

[48] ——, "Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark," in *IEEE International Symposium on Geoscience and Remote Sensing*, 2017.

[49] K. Bittner, S. Cui, and P. Reinartz, "Building extraction from remote sensing data using fully convolutional networks," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences-ISPRS Archives*, vol. 42, no. W1, pp. 481–486, 2017.

[50] A. Lagrange, B. Le Saux, A. Beaupere, A. Boulch, A. Chan-Hon-Tong, S. Herbin, H. Randrianarivo, and M. Ferecatu, "Benchmarking classification of earth-observation data: From learning explicit features to convolutional networks," in *Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International*, IEEE, 2015, pp. 4173–4176.

[51] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic segmentation of aerial images with an ensemble of cnns," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2016*, vol. 3, pp. 473–480, 2016.

[52] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," *arXiv preprint arXiv:1606.02585*, 2016.

[53] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *arXiv preprint arXiv:1606.00915*, 2016.

[54] N. Audebert, B. Le Saux, and S. Lefèvre, "Semantic segmentation of earth observation data using multimodal and multi-scale deep networks," in *Asian Conference on Computer Vision*, Springer, 2016, pp. 180–196.

[55] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[56] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 135, pp. 158–172, 2018.

[57] J. Hu, L. Mou, A. Schmitt, and X. X. Zhu, "Fusionet: A two-stream convolutional neural network for urban scene classification using polsar and hyperspectral data," in *Urban Remote Sensing Event (JURSE), 2017 Joint*, IEEE, 2017, pp. 1–4.

[58] L. Mou and X. X. Zhu, "Spatiotemporal scene interpretation of space videos via deep neural network and tracklet analysis," in *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*, IEEE, 2016, pp. 1823–1826.

[59] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint*, 2017.

[60] D. Marmanis, W. Yao, F. Adam, M. Datcu, P. Reinartz, K. Schindler, J. D. Wegner, and U. Stilla, "Artificial generation of big data for improving image classification: A generative adversarial network approach on sar data," *arXiv preprint arXiv:1711.02010*, 2017.

[61] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[62] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.

[63] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, p. 533, 1986.

[64] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[65] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26–31, 2012.

[66] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[67] H. Le and A. Borji, "What are the receptive, effective receptive, and projective fields of neurons in convolutional neural networks?" *arXiv preprint arXiv:1705.07049*, 2017.

[68] P. d'Angelo and P. Reinartz, "Semiglobal matching results on the isprs stereo matching benchmark," 2011.

[69] T. Krauß, B. Sirmacek, H. Arefi, and P. Reinartz, "Fusing stereo and multispectral data from worldview-2 for urban modeling," in *Proc. of SPIE Vol*, vol. 8390, 2012, pp. 83901X–1.

[70] R. Qin, J. Tian, and P. Reinartz, "Spatiotemporal inferences for use in building detection using series of very-high-resolution space-borne stereo images," *International Journal of Remote Sensing*, vol. 37, no. 15, pp. 3455–3476, 2016.

[71] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

Ksenia Bittner is a PhD candidate at the German Aerospace Center (DLR) since September 2015. She obtained her M.Sc. degree in Earth Oriented Space Science and Technology - an international Master's Program at the Technical University of Munich for Space Engineering and its applications: Earth System Science, Remote Sensing, and Navigation. Her interest lies broadly in the area of machine learning and its application to photogrammetry and remote sensing problems. More specifically, she is working on 3D buildings reconstruction from Digital Surface Model using deep learning techniques.

Fathalrahman Adam received the M.Sc. degree in Earth Oriented Space Science and Technology from Technical University Munich, Munich, Germany in 2014. He is currently doing his PhD research at the German Remote Sensing Data Center (DFD), one of the institutes of the German Aerospace Center (DLR). His research interests are urban detection, satellite time series analysis, and sensor fusion. He is investigating general machine learning techniques, as well as deep learning, in the context of urban remote sensing applications.

Shiyong Cui received the M.S. degree in photogrammetry and remote sensing from Chinese Academy of Surveying and Mapping, Beijing, China, in 2009. He finished the PhD at the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, Germany and obtained the Ph.D. degree in electrical engineering and computer science from Siegen University, Siegen, Germany, in 2014. His research interests include computer vision, machine learning and autonomous driving.

Marco Körner (M'15) studied Computer Sciences with Psychology as a minor subject at Friedrich Schiller University of Jena, Germany, and received his diploma (Dipl.-Inf.) and Ph.D. (Dr. rer. nat.) degrees in 2009 and 2016, respectively. From 2009 to 2015, he was a member of the Computer Vision Group in Jena. In 2012 and 2014, he was a visiting researcher at Instituto Politécnico Nacional (CIC-IPN) in Mexico City, Mexico, and the University of California, San Diego (UCSD), USA, respectively. Since 2015, he has been a senior researcher and deputy head at the chair of Remote Sensing Technology at Technical University of Munich (TUM), Germany. His main research interests focus on machine learning in computer vision, particularly for application in automotive, remote sensing, and biomedical scenarios.

Peter Reinartz (M'09) received his Diploma (Dipl.-Phys.) in theoretical physics in 1983 from the University of Munich and his PhD (Dr.-Ing) in civil engineering from the University of Hannover, in 1989. His dissertation is on optimization of classification methods for multispectral image data. Currently he is department head of the department "Photogrammetry and Image Analysis" at the German Aerospace Centre (DLR), Remote Sensing Technology Institute (IMF) and holds a professorship for computer science at the University of Osnabrueck. He has more than 30 years of experience in image processing and remote sensing and over 400 publications in these fields. His main interests are in machine learning, stereophotogrammetry and data fusion using space borne and airborne image data, generation of digital elevation models and interpretation of very high resolution data from sensors like WorldView, GeoEye, and Pleiades. He is also engaged in using remote sensing data for disaster management and using high frequency time series of airborne image data for real time image processing and for operational use in case of disasters as well as for traffic monitoring.